International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 10, Issue 1, Jan – Jun 2021; 97–118

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

OPTIMIZING DATA PIPELINES IN THE CLOUD: A CASE STUDY USING DATABRICKS AND PYSPARK

Swathi Garudasu[1], Priyank Mohan[2], Rahul Arulkumaran[3], Om Goel[4], Dr. Lalit Kumar[5] & Prof. (Dr.) Arpit Jain[6]

[1]Symbiosis Center for Distance Learning, Pune, India

[2]Scholar, Seattle University, Dwarka, New Delhi, India

[3]University at Buffalo, New York, USA

[4]ABES Engineering College Ghaziabad India

[5]Associate Professor, Department of Computer Application IILM University Greater Noida India

[6]KL University, Vijayawada, Andhra Pradesh, India

ABSTRACT

In the era of big data, organizations are increasingly reliant on cloud computing solutions to manage and process vast amounts of information efficiently. This research paper presents a case study that focuses on optimizing data pipelines using Databricks and PySpark within cloud environments. The motivation for this study stems from the growing need for organizations to enhance data processing speed, reduce operational costs, and improve resource utilization. By leveraging the capabilities of Databricks—a unified analytics platform that integrates Apache Spark with cloud technology—this research investigates the optimization strategies that can be implemented to streamline data workflows.

The case study involves the design and implementation of a data pipeline that processes a large-scale dataset. It outlines the challenges faced in traditional data processing environments, such as performance bottlenecks, high latency, and inefficient resource allocation. The paper discusses the adoption of PySpark, the Python API for Apache Spark, as a crucial tool for distributed data processing. Through the implementation of various optimization techniques—such as data partitioning, caching intermediate results, and utilizing built-in optimization tools—significant improvements in pipeline performance were achieved.

The results of the case study demonstrate notable enhancements in processing times across different stages of the data pipeline, leading to a substantial reduction in overall processing time. Furthermore, resource utilization metrics indicated improved efficiency, with lower CPU and memory usage observed post-optimization. Cost analysis also revealed a decrease in operational expenses, showcasing the financial benefits of optimizing cloud-based data workflows.

This research highlights the importance of adopting cloud technologies and modern data processing frameworks to remain competitive in today's data-driven landscape. The findings not only contribute to the field of data engineering but also provide actionable insights for organizations seeking to optimize their data pipelines. By presenting a real-world application of optimization techniques, this study serves as a valuable reference for data engineers and decision-makers aiming to enhance their data processing capabilities.

The implications of this research extend beyond the case study itself, suggesting that the methodologies employed can be adapted to various cloud environments and use cases. Future research could explore the application of these optimization strategies across different platforms and datasets, further expanding the understanding of data pipeline efficiency in cloud computing. The study concludes that embracing cloud solutions like Databricks and leveraging PySpark's capabilities can lead to significant advancements in data processing efficiency, positioning organizations to harness the full potential of their data assets.

KEYWORDS: *Databricks, PySpark, Cloud Computing, Data Pipelines, Optimization, Big Data, Scalability, Distributed Processing.*

Article History

Received: *03 Jun 2021* | ***Revised:*** *11 Jun 2021* | ***Accepted:*** *16 Jun 2021*